                  Anycast Rendevous Point (RP) mechanism using
                     Protocol Independent Multicast (PIM)
                 and Multicast Source Discovery Protocol (MSDP)

Status of this Memo

Copyright Notice

Abstract

   This document describes a mechanism to allow for an arbitrary number
   of Rendevous Points (RPs) per group in a single shared-tree Protocol
   Independent Multicast-Sparse Mode (PIM-SM) domain.

1. Introduction

   PIM-SM, as defined in RFC 2362, allows for only a single active RP
   per group, and as such the decision of optimal RP placement can
   become problematic for a multi-regional network deploying PIM-SM.

   Anycast RP relaxes an important constraint in PIM-SM, namely, that
   there can be only one group to RP mapping can be active at any time.
   The single mapping property has several implications, including
   traffic concentration, lack of scalable register decapsulation (when
   using the shared tree), slow convergence when an active RP fails,
   possible sub-optimal forwarding of multicast packets, and distant RP
   dependencies. These properties of PIM-SM have been demonstrated in
   native continental or inter-continental scale multicast deployments.
   As a result, it is clear that ISP backbones require a mechanism that
   allows definition of multiple active RPs per group in a single PIM-SM
   domain. Further, any such mechanism should also address the issues
   addressed above.

The mechanism described here is intended to address the need for
better fail-over (convergence time) and sharing of the register
decapsulation load (again, when using the shared-tree) among RPs in a
domain. It is primarily intended for applications within those
networks using MBGP, Multicast Source Discovery Protocol [MSDP] and
PIM-SM protocols, for native multicast deployment, although it is not
limited to those protocols. In particular, Anycast RP is applicable
in any PIM-SM network that also supports MSDP (MSDP is required so
that the various RPs in the domain maintain a consistent view of the
sources that are active). Note however, a domain deploying Anycast
RP is not required to run MBGP. Finally, a general requirement of
the Anycast RP scheme is that the anycast address MUST NOT be used as
the RP address in the RP's SA messages.

The keywords MUST, MUST NOT, MAY, OPTIONAL, REQUIRED, RECOMMENDED,
SHALL, SHALL NOT, SHOULD, SHOULD NOT are to be interpreted as defined
in BCP 14, RFC 2119 [RFC2119].

2. Problem Definition

The anycast RP solution provides a solution for both fast fail-over
and shared-tree load balancing among any number of active RPs in a
domain.

2.1. Traffic Concentration and Distributing Decapsulation Load Among RPs

While PIM-SM allows for multiple RPs to be defined for a given group,
only one group to RP mapping can be active at a given time. A
traditional deployment mechanism for balancing register decapsulation
load between multiple RPs covering the multicast group space is to
split up the 224.0.0.0/4 space between multiple defined RPs. This is
an acceptable solution as long as multicast traffic remains low, but
has problems as multicast traffic increases, especially because the
network operator defining group space split between RPs does not
always have a priori knowledge of traffic distribution between
groups. This can be overcome via periodic reconfigurations, but
operational considerations cause this type of solution to scale
poorly.

2.2. Sub-optimal Forwarding of Multicast Packets

When a single RP serves a given multicast group, all joins to that
group will be sent to that RP regardless of the topological distance
between the RP and the sources and receivers. Initial data will be
sent towards the RP also until configured the shortest path tree
switch threshold is reached, or the data will always be sent towards
the RP if the network is configured to always use the RP rooted
shared tree. This holds true even if all the sources and the

receivers are in any given single region, and RP is topologically
distant from the sources and the receivers. This is an artifact of
the dynamic nature of multicast group members, and of the fact that
operators may not always have a priori knowledge of the topological
placement of the group members.

Taken together, these effects can mean that (for example) although
all the sources and receivers of a given group are in Europe, they
are joining towards the RP in the USA and the data will be traversing
a relatively expensive pipe(s) twice, once to get to RP, and back
down the RP rooted tree again, creating inefficient use of expensive
resources.

2.3. Distant RP Dependencies

As outlined above, a single active RP per group may cause local
sources and receivers to become dependent on a topologically distant
RP. In addition, when multiple RPs are configured, there can be
considerable convergence delay involved in switching to the backup
RP. This delay may exist independent of the toplogical location of
the primary and backup RPs.

3. Solution

Given the problem set outlined above, a good solution would allow an
operator to configure multiple RPs per group, and distribute those
RPs in a topologically significant manner to the sources and
receivers.

3.1. Mechanisms

All the RPs serving a given group or set of groups are configured
with an identical anycast address, using a numbered interface on the
RPs (frequently a logical interface such as a loopback is used). RPs
then advertise group to RP mappings using this interface address.
This will cause group members (senders) to join (register) towards
the topologically closest RP. RPs MSDP peer with each other using an
address unique to each RP. Since the anycast address is not a unique
address (by definition), a router MUST NOT choose the anycast unicast
address as the router ID, as this can prevent peerings and/or
adjacencies from being established.

In summary then, the following steps are required:

3.1.1. Create the set of group-to-anycast-RP-address mappings

   The first step is to create the set of group-to-anycast-RP-address
   mappings to be used in the domain. Each RP participating in an
   anycast RP set must be configured with a consistent set of group to
   RP address mappings. This mapping will be used by the non-RP routers
   in the domain.

3.1.2. Configure each RP for the group range with the anycast RP address

   The next step is to configure each RP for the group range with the
   anycast RP address. If a dynamic mechanism, such as auto-RP or the
   PIMv2 bootstrap mechanism, is being used to advertise group to RP
   mappings, the anycast IP address should be used for the RP address.

3.1.3. Configure MSDP peerings between each of the anycast RPs in the
   set

   Unlike the group to RP mapping advertisements, MSDP peerings must use
   an IP address that is unique to the endpoints; that is, the MSDP
   peering endpoints MUST use a unicast rather than anycast address. A
   general guideline is to follow the addressing of the BGP peerings,
   e.g., loopbacks for iBGP peering, physical interface addresses for
   eBGP peering. Note that the anycast address MUST NOT be used as the
   RP address in SA messages (as this would case the peer-RPF check to
   fail).

3.1.4. Configure the non-RP's with the group-to-anycast-RP-address
   mappings

   Finally, each non-RP router must learn the set of group to RP
   mappings. This could be done via static configuration, auto-RP, or
   by PIMv2 bootstrap mechanism.

3.1.5. Ensure that the anycast IP address is reachable by all routers in
   the domain

   This is typically accomplished by causing each RP to inject the /32
   into the domain's IGP.

3.2. Interaction with MSDP Peer-RPF check

   Each MSDP peer receives and forwards the message away from the RP
   address in a "peer-RPF flooding" fashion. The notion of peer-RPF
   flooding is with respect to forwarding SA messages [MSDP]. The BGP
   routing tables are examined to determine which peer is the next hop
   towards the originating RP of the SA message. Such a peer is called
   an "RPF peer". See [MSDP] for details of the Peer-RPF check.

3.3. State Implications

   It should be noted that using MSDP in this way forces the creation of
   (S,G) state along the path from the receiver to the source. This
   state may not be present if a single RP was used and receivers were
   forced to stay on the shared tree.

4. Security considerations

   Since the solution described here makes heavy use of anycast
   addressing, care must be taken to avoid spoofing. In particular
   unicast routing and PIM RPs must be protected.

4.1. Unicast Routing

   Both internal and external unicast routing can be weakly protected
   with keyed MD5 [RFC1828], as implemented in an internal protocol such
   as OSPF [RFC2328] or in BGP [RFC2385]. More generally, IPSEC
   [RFC2401] could be used to provide protocol integrity for the unicast
   routing system.

4.1.1. Effects of Unicast Routing Instability

   While not a security issue, it is worth noting that if unicast
   routing is unstable, then the actual RP that source or receiver is
   using will be subject to the same instability.

4.2. Multicast Protocol Integrity

   The mechanisms described in [RFC2362] should be used to provide
   protocol message integrity protection and group-wise message origin
   authentication.

4.3. MSDP Peer Integrity

   As is the the case for BGP, MSDP peers can be protected using keyed
   MD5 [RFC1828].

5. Acknowledgments

   John Meylor, Bill Fenner, Dave Thaler and Tom Pusateri provided
   insightful comments on earlier versions for this idea.

   This memo is a product of the MBONE Deployment Working Group (MBONED)
   in the Operations and Management Area of the Internet Engineering
   Task Force. Submit comments to <mboned@ns.uoregon.edu> or the
   authors.

6. References

   [MSDP]   D. Meyer and B. Fenner, Editors, "Multicast Source
               Discovery Protocol (MSDP)", Work in Progress.

   [RFC2401]   Kent, S. and R. Atkinson, "Security Architecture for the
               Internet Protocol", RFC 2401, August 1995.

   [RFC1828]   Metzger, P. and W. Simpson, "IP Authentication using Keyed
               MD5", RFC 1828, August 1995.

   [RFC2119]   Bradner, S., "Key words for use in RFCs to Indicate
               Requirement Levels", BCP 14, RFC 2119, March 1997.

   [RFC2362]   Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering,
               S., Handley, M., Jacobson, V., Liu, C., Sharma, P. and L.
               Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM):
               Protocol Specification", RFC 2362, June 1998.

   [RFC2328]   Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

   [RFC2385]   Heffernan, A., "Protection of BGP Sessions via the TCP MD5
               Signature Option", RFC 2385, August 1998.

   [RFC2403]   Madson, C. and R. Glenn, "The Use of HMAC-MD5-96 within
               ESP and AH", RFC 2403, November 1998.

7. Author's Address

   Dorian Kim
   Verio, Inc.
   EMail: dorian@blackrose.org

   Hank Kilmer
   EMail: hank@rem.com

   Dino Farinacci
   Procket Networks
   EMail: dino@procket.com

   David Meyer
   EMail: dmm@maoz.com

8. Full Copyright Statement

Acknowledgement